ACED: Large Language Model-based Research Agent for Synthetic Biology Research Idea Generation

Overview

Scientific research drives innovation, advances knowledge, and improves lives. Synthetic biology, in particular, can enable biomanufacturing, personalized medicine and advanced therapeutics. The biological research process involves formulating and validating new ideas through comprehensive experiments, which require significant effort in extracting and synthesizing knowledge from vast scientific literature and new data. Designing valid experiments also demands experience and extensive discussions. These challenges highlight the need for automated tools to distill knowledge and generate research ideas and plans.

Inspired by the enormous success of large language models (LLM) in many domains, we propose developing LLM-based framework to automate the process. Our work will improve the automation of synthetic biology research idea generation, coupled with fundamental advances in large language models, by building a virtuous cycle involving the two. We identify multiple key challenges: 1) General domain LLM lacks knowledge and the understanding of diverse biology tokens/data (e.g. genomic sequence); and principles/laws of biological domain (e.g. genetic regulation and expression, mass action law). 2) There are vast amount of existing literature to draw insights from when prompting LLM to generate new ideas, directly inputting all of them to LLM is inefficient and not feasible. 3) Raw research ideas generated in a single step often lack the necessary refinement of various aspects (including experimental feasibility). This can lead to several issues, such as triviality and lack of novelty. In general, challenge 1 requires collecting data and borrowing insights from the biology domain to **achieve fundamental advances in LLMs**; challenge 2 and 3 require building **an advanced LLM-based framework to enhance scientific discovery** in synthetic biology.

We will overcome these challenges in following research goals (RGs). RG1 conducts data collection and synthetic data generation in the synthetic biology domain. We will i) select and process scientific papers/results from existing knowledge base, focusing on the diversity of synthetic biology data; ii) collect domain-specific scientific principles that all phenomena should adhere to. RG2 improves the LLM by novel evolution-inspired pre-training and automatic LLM merging strategy. RG2 then fine-tune LLM with synthetic data generated by biology principle-based simulations to improve its faithfulness. Together, RG1 and RG2 tackle challenge 1 by leveraging established knowledge/principles in biology to develop better LLMs. RG3 tackles the challenge of extracting key information from the vast amount of literature and leveraging it for proposing candidate hypotheses. RG4 refines the generated candidate hypotheses with conceptual and empirical feedback. **Our approach establishes a virtuous cycle of innovation where advancements in LLM technology enhance the ability to generate novel synthetic biology ideas, and the validated hypotheses provide new data and insights to improve the LLMs**. This cycle not only accelerates scientific discovery in synthetic biology but also drives continuous improvements in AI methodologies.

Keywords: Synthetic Biology, Information Extraction, Hypotheses Generation, Large Language Models.

Intellectual Merit

The proposal will (1) introduce novel LLM merging methodologies inspired by evolution theory in biology, generalizable beyond the biology domain and applicable to other tasks. (2) advance the state-of-the-art and more faithful LLM for domain-specific scientific literature understanding applications, such as hypotheses generation and experimental plan generation; (3) produce tools for generating better research ideas for scientific research in synthetic biology; (4) develop new data collection and synthetic data generation strategies that will provide community-wide resources to the public for advancing research in related areas.

Broader Impacts

Our proposed work will include at least the following broad impacts: (1) our idea generation framework will expedite the research process in the synthetic biology domain, accelerating the development of techniques such as genome editing. Additionally, it will be applicable to other scientific domains. (2) the knowledge extraction system can be utilized for building knowledge graphs from scientific documents/literature (e.g. biomedical events extraction). (3) we will contribute to education by mentoring graduate, undergraduate, and K-12 students, with a particular emphasis on supporting female students. Tools developed based on our methods will help provoke critical thinking.

ACED: Large Language Model-based Research Agent for Synthetic Biology Research Idea Generation

1 Introduction

Synthetic biology stands at the forefront of scientific innovation, offering transformative possibilities in various domains, including personalized medicine and advanced therapeutics. By leveraging principles of biology, engineering, and computational science, synthetic biology enables the design and construction of new biological parts, devices, and systems, as well as the re-design of existing, natural biological systems [4, 25, 44, 76]. This multidisciplinary field has already demonstrated its potential to revolutionize areas such as genetic engineering and metabolic pathway optimization [72]. Despite its promising potential, the research process in synthetic biology remains complex and labor-intensive, often requiring the synthesis of vast amounts of scientific literature and new experimental data [34, 82]. Experienced researchers must meticulously extract and integrate knowledge, formulate new research hypotheses, and design valid experimental frameworks, necessitating extensive expertise and collaborative discussions. The increasing volume of published research exacerbates these challenges, making it difficult for scientists to stay up-todate with the latest advancements and to identify gaps in current knowledge [60]. For example, the number of academic papers published per year is over 7 million [26]. These limitations underscore the pressing need for advanced automated tools that can assist researchers in distilling knowledge from large datasets and generating innovative research ideas (i.e. hypotheses generation and experimental design).

Recently, Large Language Models (LLMs) have shown impressive capabilities in generating text with remarkable quality [39, 68, 81] across diverse specialized domains including math, physics, biology and medicine. Thus, LLMs may be a transformative tool to accelerate the scientific research process. Specifically, LLMs can process and analyze large volumes of data at a speed and scale that is beyond human capabilities. They can also identify patterns, trends, and correlations that may not be immediately apparent to human researchers, potentially uncovering novel research opportunities that might otherwise go unnoticed.

Motivated by the need for automated idea generation tools and the success of LLMs, **we propose the first Large Language Model-based Research Agent (LLM-RA), developed to assist synthetic biology researchers by efficiently generating feasible and novel research idea hypotheses and experiment design from diverse and complex types of scientific literature data**. It should enable researchers to stay current with the latest advancements and identify novel ideas more rapidly. We identify the following key methodological challenges.

First, general domain LLM such as GPT-4 [68, 69] lacks knowledge in the biology domain, especially the understanding of diverse tokens/language (e.g. genomic sequence, protein structure) [6, 49]. In addition, the model's predictions are not guaranteed to adhere to established principles of biology, related to regulation and conservation of energy and matter. For example, they generate unfaithful contents ("hallucinations") that cannot be grounded to domain-specific scientific facts [37], such as claiming "SyntheticBioX enzyme can convert any type of waste into biofuel with 100% efficiency", which is scientifically inaccurate and not supported by known principles of synthetic biology.

Second, current work focuses on the narrow idea generation scope: identifying new relationships between two fixed concepts (one paper and its few references) [73, 83, 92]. We focus on building an LLM-powered agent that is capable of generating research ideas over *vast amount of scientific literature*. The major challenge is the accumulated contextual knowledge over numerous papers, which skilled human researchers either possess or learn from perusals of scientific literature, then leverage to come up with and develop new ideas [5]. However, current LLMs are not able to input and process such a large amount of information regarding existing literature to generate research ideas, due to its limited context window [68, 70].

Third, the traditional one-step generation approach (concludes once the ideas are formulated) lacks of an iterative refinement process based on reviews and feedback from multiple perspectives (e.g. clarity, novelty, experimental feasibility) [92], which differs from typical human-driven research processes, which develop and improve research ideas through multiple rounds of reflections and multiperspective feedback (e.g. conceptual and empirical).

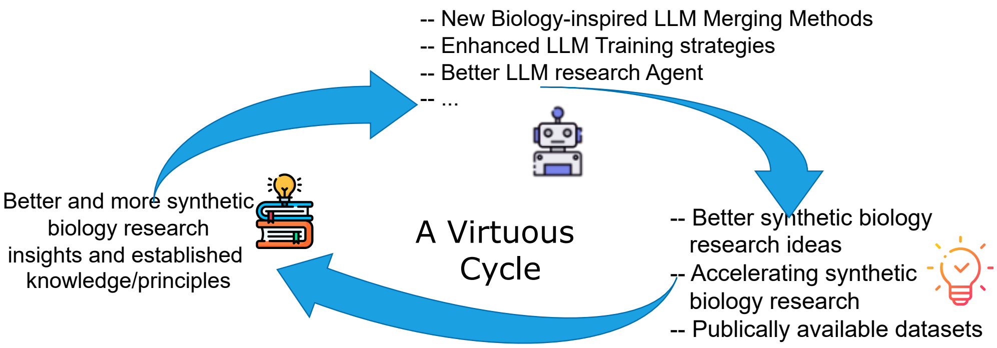

-- New Biology-inspired LLM Merging Methods
-- Enhanced LLM Training strategies
-- Better LLM research Agent
-- ...
Better and more synthetic biology research insights and established knowledge/principles
A Virtuous Cycle
-- Better synthetic biology research ideas
-- Accelerating synthetic biology research
-- Publically available datasets

Figure 1: The virtuous cycle of our innovations leading to both scientific and computational advances.

1.1 Proposed Research

To sum up, to automate the idea generation process, we need to tackle the challenges of developing LLMs that can handle diverse types of data in the biology domain and adhere to scientific principles; and LLM-powered agent that can i) leverage broad research literature to come up with research ideas; ii) conduct iterative refinement based on comprehensive feedback (conceptual and empirical). Finally, to benchmark methods' performance in generating research ideas, we need to design novel and fair evaluation methods that align with real research values. PIs aim to tackle the challenges by investigating the following research goals (RGs) with novel methodologies, based on their unique qualifications in both LLMs (PI Du) and biological science (PI Bleris). These research goals contribute to building core components of the pipeline for generating research ideas in Figure 2, which will be utilized by real synthetic biology researchers.

RG1: Synthetic Biology Domain Knowledge and Dataset Collection. RG1 first formalizes the task, then conducts data collection and synthetic data generation in the synthetic biology research domain. For collection, we will i) select and process scientific papers/results from the existing knowledge base (e.g. PubMed), focusing on the diversity of synthetic biology data. We will also consider adding LLM-generated and validated hypotheses as new knowledge, facilitated by the technique proposed in this proposal; ii) collect domain-specific scientific principles that all predictions should adhere to (e.g., gene expression regulation).

RG2: Biology-inspired Training of Large Language Models. RG2 enhances the LLM using evolution-inspired pre-training and merging strategies to improve faithfulness and reduce hallucinations. It then fine-tunes the LLM with synthetic data from biology principle-based simulations. We draw insights from our experience in training large vision language models (LVLMs) [40, 42]. To close the cycle, we will augment this knowledge data with hypotheses/data generated from RG4 (Figure 1). Together, RG1 and RG2 tackle the challenge of LLM's lack of knowledge by training with large amounts of data in the synthetic biology domain, as well as leveraging insights in biology to develop better LLMs.

RG3: LLM agent for Hypotheses Proposal and Experiment Design by Leveraging Knowledge Graph (KG). We aim to build an LLM-powered research agent capable of generating research ideas from the entire body of scientific literature, unlike existing methods that consider only 1-2 publications [83, 92]. We will construct a KG (leftmost of Figure 2) with entity-relation information from scientific articles to enhance hypothesis generation and experiment design. To address the limited coverage in traditional information extraction systems, we build on top of our work in information extraction [9, 16, 17, 18, 19, 20, 22] and question generation [14, 15, 21], to propose a novel question-answer pair paradigm to represent entity-relation pairs.

RG4: Refining Hypotheses with Conceptual and Experimental Feedback. Candidate hypotheses generated by LLMs are often of low quality and infeasible [93]. To address this, based on our prior work on inductive reasoning [28, 92, 93, 94], we propose leveraging conceptual feedback (e.g., validity, novelty) and experimental feedback to refine these hypotheses (bottom right of Figure 2). We will augment the refined/validated hypotheses to RG2 for training better LLMs. We will design comprehensive, reference-free methods for **evaluation of research ideas**, which will be prompting-based and consider multifaceted

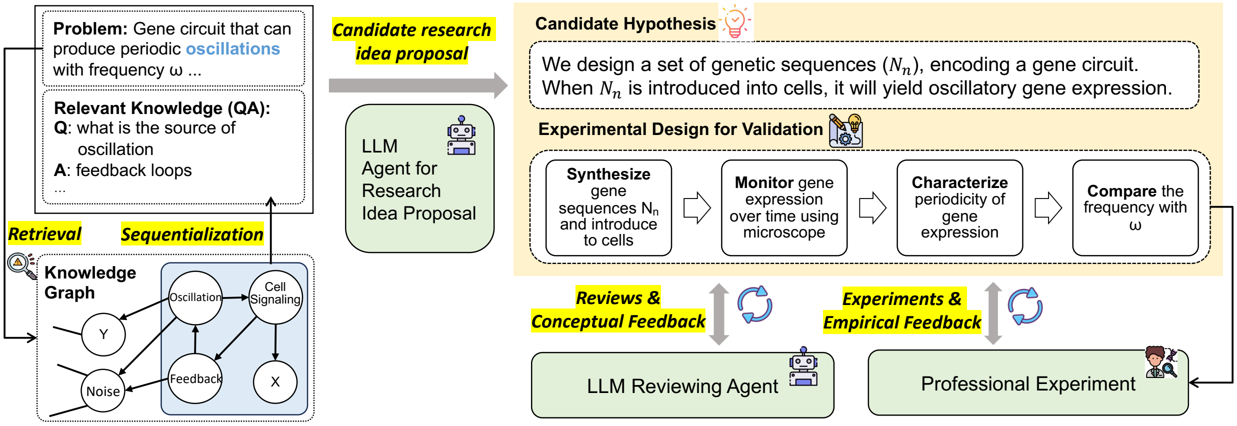

Problem: Gene circuit that can produce periodic oscillations with frequency ω ...
Relevant Knowledge (QA):
Q: what is the source of oscillation
A: feedback loops
...
Retrieval
Sequentialization
Knowledge Graph
Oscillation
Cell Signaling
Y
Noise
Feedback
X
Candidate research idea proposal
LLM Agent for Research Idea Proposal
Candidate Hypothesis
We design a set of genetic sequences (N_n), encoding a gene circuit. When N_n is introduced into cells, it will yield oscillatory gene expression.
Experimental Design for Validation
Synthesize gene sequences N_n and introduce to cells
Monitor gene expression over time using microscope
Characterize periodicity of gene expression
Compare the frequency with ω
Reviews & Conceptual Feedback
Experiments & Empirical Feedback
LLM Reviewing Agent
Professional Experiment

Figure 2: Our framework for Large language Model-based Research Idea Generation.

aspects of the ideas.

As indicated in Figure 1, **our approach establishes a virtuous cycle of innovation where advancements in LLM technology enhance the ability to generate novel synthetic biology ideas, and the validated hypotheses provide new data and insights for inventing novel groundbreaking methodology for improving the LLMs**. This cycle not only accelerates scientific discovery in synthetic biology but also drives continuous improvements in AI methodologies, fostering interdisciplinary collaboration and innovation.

1.2 Intellectual Merit

Taken together, the fundamental training and merging methods and datasets developed in the proposed research will support building powerful LLMs with an improved understanding of the synthetic biology domain. Based on it, we introduce the LLM agent-based research idea generation framework, which will significantly expedite and contribute to the research in the synthetic biology. Our proposed research will: (1) advance the state-of-the-art of scientific (and general) large language models, by designing better training and model merging strategies inspired by biology principles and the collected domain-specific dataset; (2) develop novel techniques for constructing knowledge graphs for scientific literature; Moreover, our project will (3) develop new data collection and synthetic data generation strategies that will provide community-wide resources to the public for advancing research in related areas; and (4) introduce reference-free evaluation strategies for comprehensive aspects including hypotheses novelty and feasibility.

2 Broader Impacts

Education and Dissemination of Results. The Bioengineering Department at the University of Texas at Dallas offers undergraduate and graduate degrees since 2011. The 2014 freshmen year leads in numbers in the Engineering School, with a large portion of our students pursuing the pre-med track. More generally, the University has among the best-prepared incoming freshmen each year in the state of Texas. In recent years, nearly 40% of UT Dallas freshmen ranked in the top 10 percent of their high school class and 75% ranked in the top 25 percent. PI Bleris will continue to offer summer research opportunities to undergraduate students and has developed graduate-level courses on “Systems Biology” and “Genome Editing.” Students at the graduate level can be exposed to a carefully planned and broad spectrum of topics from diverse scientific disciplines. These courses will be expanded to include lectures intersecting machine learning, large language models, and molecular biology, integrating the results of the proposed research. Students will use the research idea generation tool for their course reading materials and provide feedback on its strengths and weaknesses, learning to propose new circuits and gene sequences.

PI Du will train graduate and undergraduate students in NLP and AI, involving them in project meetings, seminars, and conferences to provide a stimulating educational experience. He will provide hands-on research experience and exposure to interdisciplinary studies. Students will participate in project meetings, seminars, and academic conferences, gaining valuable insights and practical knowledge.

Together, we propose to develop **a customized, special topics class on LLM and synthetic biology**, which will be offered to students from both CS and Biology. The course will help students understand the interdisciplinary knowledge and get interested in the general area of AI for Science [2, 7]. **Evaluation.** During the courses, PIs ask students to write 2-3 sentences comments at the end of each assignment submission. At the end semester, UTD provides student evaluations to measure course quality. PI will be collaborating with the UTD Center for Teaching and Learning (CTL) to evaluate teaching and get feedback.

In addition, we will collaborate with the CTL to develop **activities on campus** that allow students to use our proposed research idea generation tools in critical thinking exercises on topics of their choice. Specifically, the exercises will be developed to help inspire students on future work or extensions they can make upon what they have learned in class. The participants will be selected from the "UTD Program for High School Students 2024" [1], which will be held between May and August 2025. Generally, the learning process will introduce students to the basic AI/NLP/Biology topics.

Promoting Diversity in STEM Education. Both PIs are committed to promoting diversity in STEM. We will contribute to education by mentoring graduate, undergraduate, and K-12 students, with a particular emphasis on supporting women students. PI Bleris will broaden the participation of underrepresented minorities through the "Academic Bridge program at UT Dallas" and continue to support female students in STEM. PI Du is dedicated to maintaining a diverse research group, where one-third of his PhD students are women. He has been the faculty consulting member for the UT Dallas on-campus activities such as Girls Who Code and UTD Society of Women Engineers (SWE). He will keep recruiting PhD and Master students from the under-represented groups as mentioned above. **Evaluation.** PI will record how many students from minority groups they recruit and will also monitor their outcomes.

Data Resources and Tools. The project will produce valuable data resources for researchers, including: A scientific hypothesis discovery dataset containing prompts for designing gene circuits based on different requirements, along with expert-written hypotheses; A comprehensive knowledge graph dataset for specific synthetic biology literature in the question-answer pair format; A dataset with open-ended prompts to test LLM's understanding of scientific knowledge/principles in the synthetic biology domain. These tools developed based on our work will support research across multiple scientific disciplines, enhancing data accessibility and utility.

3 Contributions to Computing and Scientific Discovery

The project aims to make breakthroughs in the fields of computing and synthetic biology, through the development of advanced large language models (LLMs) specifically tailored for scientific research. The novel contributions to both computing and scientific disciplines are outlined below.

3.1 Contributions to Computing

Novel LLM Merging Methodologies. Inspired by evolutionary principles, we introduce innovative model merging techniques to automate the creation of powerful foundation models. These techniques enhance the integration of topic/domain-specific knowledge into a single LLM, leveraging their collective intelligence without requiring extensive additional training data or computational resources. This approach can be applied in any setting where diverse LLMs exist, allowing for the automatic creation of more powerful models. In general, this work contributes new state-of-the-art models to the open-source community and introduces a new paradigm for automated model composition.

Enhanced Training and Refinement Strategies. We introduce a novel fine-tuning with the fine-grained feedback framework that can incorporate any scientific principles into AI reasoning mechanisms, mitigating the problem of hallucinations and increasing the factual correctness and trustworthiness of any LLMs. In addition, we propose a simulation-based synthetic generation approach for augmenting training data. Regarding the prompting-based refinement strategy, we propose a novel iterative hypothesis refinement process. This strategy ensures that the generated content is not only innovative but also satisfies other constraints, which can be applied to improve any text-writing tasks in general.

Knowledge Graph Construction. We propose advanced techniques for extracting and constructing knowledge graphs from scientific literature, leveraging question-answer pair paradigms to represent entity-relation pairs. This paradigm enhances the model's ability to understand complex relationships within the data and extract the entirety of entity-relation triples. It benefits NLP areas like information retrieval, semantic search, and automated reasoning across multiple scientific disciplines. Our sequentilization strategy for leveraging KG can also be applicable to other tasks in NLP such as question answering.

3.2 Contributions to Scientific Discovery

Accelerating Synthetic Biology Research. Our LLM-powered research agent expedites the generation of novel and feasible research hypotheses and experimental designs, significantly reducing the time and effort required by researchers. This acceleration fosters rapid advancements in synthetic biology, enabling breakthroughs in personalized medicine, genetic engineering, and metabolic pathway optimization.

Integrating Biological Principles. By embedding principles such as genetic regulation, expression, and conservation of mass into the LLM training process, our approach ensures that generated hypotheses are biologically plausible and scientifically sound. This integration bridges the gap between computational models and real-world biological systems.

Publicly Available Resources. The datasets, knowledge graphs, synthetic data generation strategies, and evaluation methods developed through this project will be made publicly available, providing valuable resources for the broader scientific community. This openness facilitates further research and collaboration across various scientific fields.

3.3 Generalization to Other Scientific Domains

The methodologies and frameworks developed in this project have the potential to generalize to a wide range of scientific disciplines beyond synthetic biology. The LLM merging techniques, scientific knowledge-guided fine-tuning strategies, and knowledge graph construction methods can be adapted to fields such as biomedical research, environmental science, materials science, etc. By ensuring that the models are grounded in domain-specific knowledge and capable of processing vast amounts of literature, our approach can accelerate scientific discovery across numerous domains. For example, as indicated by our preliminary study, the iterative refinement mechanism can also contribute to hypotheses generation in the computational social science domain [93].

3.4 Collaborative Impact

The interdisciplinary expertise of our team, combining strengths in computing (PI Du) and biological science (PI Bleris), ensures that the collaborative contributions of this project will exceed the sum of individual efforts. By integrating cutting-edge computational techniques with deep synthetic biological insights, we create a powerful synergy that drives innovation and advances in both fields.

4 RG1: Preparation of Synthetic Biology Task, Pre-training Data, and Domain Knowledge Collection

Task Setup and Preparation. Synthetic genetic circuits are engineered DNA sequences integrated into the cell's genome. They leverage the cell's machinery to function as designed, processing inputs, and producing outputs, thereby enabling new functionalities and behaviors within the cell.

A synthetic genetic circuit operates as an information processing system within the cellular environment [23, 77, 90], capable of responding to various biological events and triggering significant changes in the molecular composition of the cell. This ability to process internal biological signals has numerous applications in basic research and medicine. Our previous work has demonstrated the successful operation of such gene circuits in human cells [23, 27, 44, 50, 75, 76, 77, 89]. In this task, **our goal is to build a novel framework for the rapid assembly, stable integration, and comprehensive characterization of synthetic genetic circuits in mammalian cells**. Utilizing Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) components as modules within our synthetic networks, our approach aims to significantly advance the mammalian synthetic biology field by overcoming slow experimental timescales and facilitating future research.

For a certain problem, given the vast parameter space and the rich theoretical and empirical knowledge in the scientific literature on gene regulation, an LLM-based agent can facilitate the generation of a hypothesis including specific genetic sequences that encode genetic circuits with desired gene expression profiles. For instance, when tasked with creating a toggle switch, the LLM will suggest implementations involving two mutually inhibiting genes, specifying thresholds and binding strengths, along with the corresponding genetic sequences that encode the necessary proteins and RNA. The LLM agent will be built based on Research Goals 2,3, and 4.

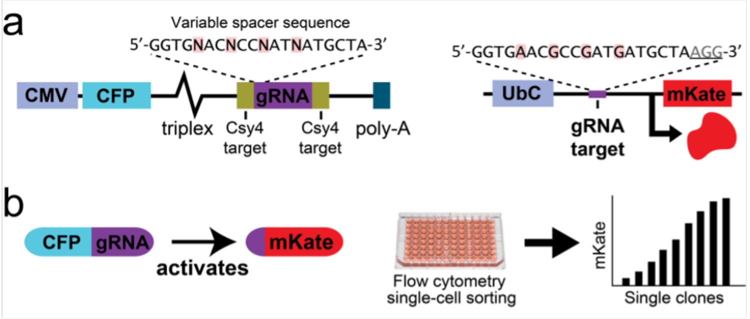


Figure 3: **Library of sgRNAs**. (a) A sgRNA library with a variable region targeting a specific target upstream from a fluorescent reporter. (b) The sgRNA library will be delivered in cells and individual cells will be sorted and expanded. Each clone will harbor a unique sgRNA and therefore will have different strength of inhibition, resulting in a massively parallel scan of all negative inhibition strengths.

RG1.1: Pre-training Data Collection. The data collection task will involve gathering a comprehensive set of literature and experimental results required to pre-train the LLM (functioning as the assembly tool) for synthetic genetic circuits. This includes: 1) **Literature and Database Integration**: Incorporating data from scientific literature, genomic databases (such as ENCODE and GEO), and previous experimental results to provide a rich dataset for pre-training. This will include gene expression profiles, regulatory sequence information, and known interactions in synthetic biology. 2) **sgRNA Library** (Figure 3): Preparing a sgRNA library with variable regions targeting specific sequences upstream from a fluorescent reporter (mKate). This library will enable a wide spectrum of connection strengths and allow for the retrieval of clones with different affinities. 3) **Stable Integration**: Incorporating circuits into safe harbor loci within mammalian cells to ensure controlled transgene copy numbers and locations; 4) **Dynamic Analysis**: Analyzing the dynamics of synthetic circuits through time-lapse single-cell measurements. Pre- and post-perturbation states will be quantified at single-molecule mRNA and single-cell protein levels to correlate mRNA to protein levels for selected architectures;

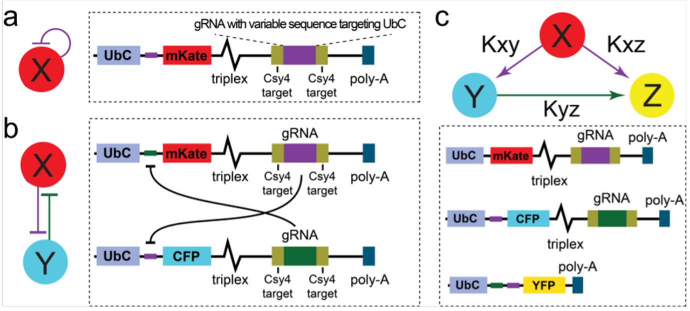


Figure 4: **Examples of circuit architectures**. The sgRNA library will be delivered in cells and individual cells will be sorted and expanded. Each clone will harbor a unique sgRNA and therefore will have different strength of inhibition, resulting in a massively parallel scan of all negative inhibition strengths. These include (a) negative feedback, (b) two-node toggle, and (c) three-node circuits.

RG1.2: Domain Background Knowledge Collection. Domain knowledge collection involves gathering established facts, constraints, and methodologies within the discipline of synthetic biology for scientific principle-guided fine-tuning of the LLM, including: 1) **Synthetic Genetic Circuit Design**: Detailed information on designing synthetic genetic circuits and CRISPR technology (Figure 4); 2) **Quantitative Rules in (Synthetic) Biology**: Incorporating specific quantitative models and equations relevant to synthetic biology, such as Michaelis-Menten kinetics for enzyme reactions, Hill equations for cooperative binding, and conservation laws in metabolic networks. These models are directly related to the properties and behaviors of generated circuits, ensuring that the LLM's predictions about circuit dynamics, efficiencies, and interactions are grounded in well-established scientific principles. 3) **Genome Editing Limits**: Exploring the limits of editing multiple genetic loci and integrating large DNA fragments in various cell lines, such as HEK293, HeLa, and CHO cells; 4) **Previous Research Findings (including findings generated by ourselves in the project**: Incorporating findings from previous research papers, including validated hypotheses and experimental results, into the knowledge base to continuously update and refine the LLM's understanding. 5) **Experimental Techniques**: Techniques for tracking individual mRNA in cells, such as PrimeFlow technology, which allows rapid assessment of mRNA distributions under different conditions.

This knowledge will support the development of master cell lines with inducible expression of components like dCas9, MCP-VP64 (activator), PCP-KRAB (repressor), and Csy4. Using these cell lines will permit rapid design and testing of scalable network architectures using custom sgRNAs proposed by the LLM. Additionally, Csy4 will be integrated to fine-tune sgRNA expression within the UTR of a gene.

5 RG2: Biology-inspired Training of Large Language Models

Overview and Motivations. In this research goal, we develop the base large language models (LLM) specializing in understanding data from the synthetic biology domain. It will be the foundation model used in the framework for the proposition (Sec. 6) and refinement (Sec. 7) of the research ideas including the hypotheses. There are **two major challenges**: (1) Open-source and general domain LLMs such as LlaMA [39, 81, 96] are trained on web corpus with natural language and contain little knowledge of specific domains such as biology, especially more fine-grained direction such as synthetic biology. There are diverse types of "language" used in biology such as genetic sequences (i.e. "ATGC"), protein structures (e.g. amino acid), and general literature text (e.g. BRCA1 gene, metabolic pathway); (2) There is no guarantee that the model's predictions adhere to scientific and biological facts/rules [12, 37], especially considering the complexity of language used in a specific domain of synthetic biology. Below in RG2.1 we first describe how to pre-train the general purpose base LLM for each type of data, and how to merge them for stronger models. Then in RG2.2, we design a fine-tuning framework to enhance LLM's reasoning capabilities regarding scientific facts, constraints, and principles in synthetic biology.

5.1 RG2.1: Evolutionary Model Training and Merging

Models Pre-training on Diverse Types of Data. First, we will conduct pre-training of models with instructions data [70] from diverse topics that are closely relevant to synthetic biology, including gene prediction, protein function prediction and general literature mining. We pre-train on top of the state-of-the-art open-source LLM Mistral [39] to obtain models specified in each type of language. More specifically, we will build three LLMs for understanding: 1) Literature and Database Integration (LLM_{lit}); 2) DNA and sgRNA (LLM_{DNA}); and 3) Circuit Dynamics and Integration ($LLM_{circuit}$). Each of them will be trained on relevant data collected in RG1.1.

For LLM_{lit} for general literature mining, example tokens include Scientific Terms: Keywords and phrases from literature such as gene names ("BRCA1 gene"), and scientific terms ("metabolic pathway"). For LLL_{DNA}, we will train it with data generated from the sgRNA library. Example tokens include Nucleotides: Single letters representing DNA bases (A, T, C, G). Although there are existing models such as DNA-BERT [36] and HyenaDNA [65] for modeling DNA sequences, they are not for our specific problem of synthetic biology, and they are based on BERT [13] instead of the stronger Mistral, which we use in LLM merging (below). $LLM_{circuit}$ for Circuit Dynamics and Integration will be trained with synthetic circuit design data (e.g. "integrate construct1 into safe harbor locus AAVS1 in HEK293 cells") and dynamic analysis techniques (e.g. "mRNA levels at 50 copies per cell"). Similarly, we can train LLM with protein function prediction data, example tokens include Amino Acids (e.g. "MVLTIYPDELVQIV..."), but we will leave it as an optional choice.

Automated LLM Model Merging based on Evolution Algorithm [87] demonstrated that merging can be used to improve robustness to domain shift in fine-tuned models by averaging the parameters of the original pre-trained model with the fine-tuned parameters. Also, it is a cost-effective approach for developing new models. We propose merging our pre-trained LLMs which specialize in different types of data (biology language), into a single architecture. Apart from our pre-trained models, we exploit other open-source models from HuggingFace [85] and add them into our candidate set of LLMs to be merged (e.g. BioMistral [48], ProtBERT [24], BioBERT [49]). Since we choose the base LLM as Mistral [39] as the base model previously when training with our collected data ($LLM_{circuit}$, LLM_{lit}, LLM_{DNA}). We use BioMistral and the general domain Mistral as the additional LLMs.

However, LLM merging (e.g. averaging parameters) traditionally relies on much human intuition, many trials and domain knowledge, limiting its potential [61, 79, 86]. Especially, if we consider many of the LLMs, the exploration space would be too large and human intuition can only go so far. **Our goal is to create a unified framework capable of automatically generating a merged model from a selection of foundation models, ensuring that the performance of this merged model surpasses that of any individual model in the collection**. Drawing inspiration from biological research and complex logic circuits [89, 90], we propose the Evolutionary Model Merge, an innovative method that leverages evolutionary algorithms to optimize the model merging process. To systematically address this challenge, we first decompose the merging process into two distinct, orthogonal configuration spaces: Data Flow Space (Layers) (Figure 6) and Parameter Space (Weights) (Figure 5). By analyzing their individual impacts, we can refine the merging intricacies. Building on this analysis, we introduce a cohesive framework that seamlessly integrates these spaces. In the two Figures, "block" means the input/output embedding layers or a transformer layer.

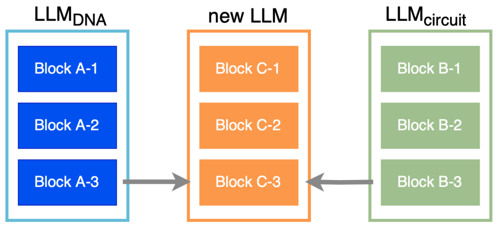


Figure 5: Merging Models in the Parameter Space (Weights).

Biological systems continuously undergo genetic mutations, and natural selection optimizes these mutations to improve the fitness of organisms. Beneficial mutations are retained and propagated, leading to more efficient and optimized organisms. We design **Parameter Space Merging** (Figure 5), model parameters are optimized to enhance performance. This process can be likened to the natural selection of beneficial traits, where the best-performing parameter configurations are identified and merged to improve the overall model efficiency and accuracy.

Model merging in the PS aims to integrate the weights of multiple foundational models into a unified entity with the same neural network architecture, yet outperforming the individual models. Following [3], we leverage task vectors analysis to understand each model's strengths, based on the specific tasks they are optimized for or excel in [35]. We establish merging configuration parameters for sparsification and weight mixing at each layer, including input and output embeddings. These configurations are then optimized using an **evolutionary algorithm, such as CMA-ES [30]**, for selected tasks, guided by accuracy on ProteinGLUE [8] (a benchmark for measuring LLM's capability in biomedical tasks).

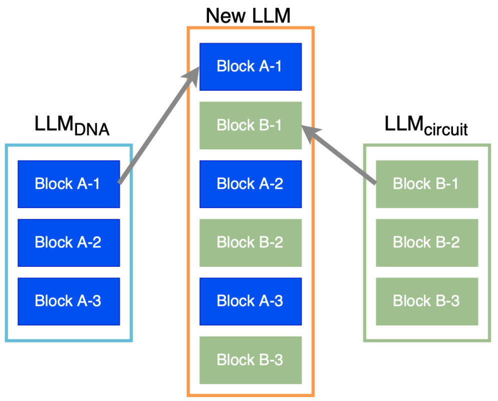


Figure 6: Merging Models in the Data Flow Space (Layers).

In biology, genetic diversity and the variety of evolutionary pathways ensure that species can adapt to changing environments. Different genes and traits are expressed depending on environmental stimuli and survival needs. Inspired by this, we design **Data Flow Space (Layers) Merging** (Figure 6), multiple data pathways can be optimized and merged to ensure that the most relevant and informative data flows are utilized during model inference. This approach mimics the adaptability and robustness found in natural ecosystems, where diverse pathways lead to more resilient outcomes

More specifically, DFS uses evolution to discover the best combinations of the layers of LLMs to form a new model. Unlike merging in PS, merging in DFS preserves the original weights of each layer unchanged. Instead, it optimizes the inference path that tokens follow as they traverse through the neural network. For example, after the *i*-th layer in model *A*, a token may be directed to the *j*-th layer in model *B*. Following [3], we adopt CMA-ES in EvoJAX [80] and optimize the hyperparameters for a total of 100 generations.

Evaluations. We will test of final LLM on ProteinGLUE [8] benchmark: a set of seven per-amino-acid tasks for evaluating learned biomedical elements representations. **To investigate if our LLM merging strategies**

could generalize to other scientific domains, we will conduct evolutionary merging with LLMs from multiple disciplines (i.e. biology, chemistry, Math) and test the single models performance on benchmarks of various domains (e.g. GSM8K [11] for Math, MMLU [33] for chemistry and biology), we will compare with the state-of-the-art performance on each benchmark.

5.2 RG2.2: Biology Knowledge-guided Fine-tuning

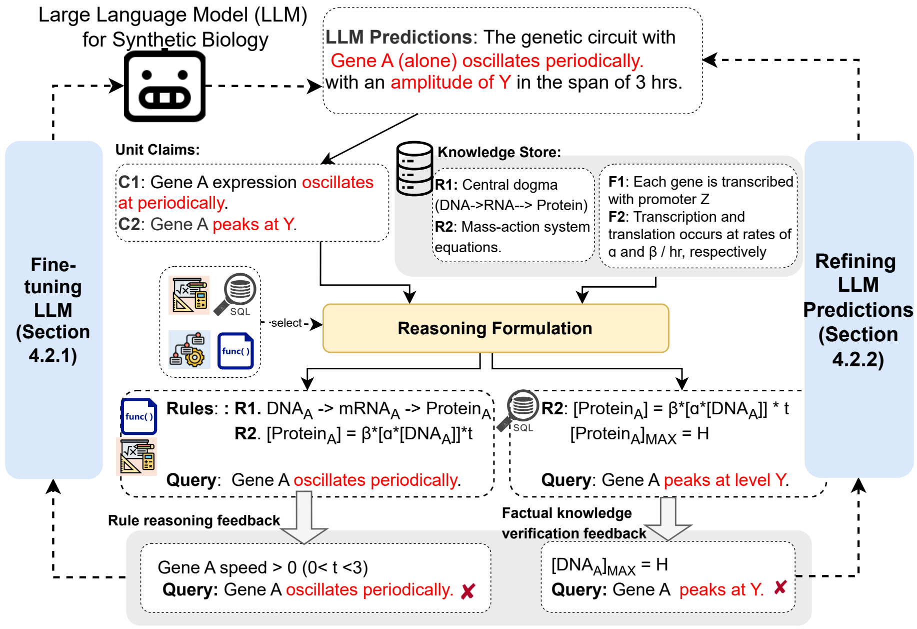

Large Language Model (LLM) for Synthetic Biology
LLM Predictions: The genetic circuit with Gene A (alone) oscillates periodically. with an amplitude of Y in the span of 3 hrs.
Unit Claims:
C1: Gene A expression oscillates at periodically.
C2: Gene A peaks at Y.
Knowledge Store:
R1: Central dogma (DNA->RNA--> Protein)
R2: Mass-action system equations.
F1: Each gene is transcribed with promoter Z
F2: Transcription and translation occurs at rates of α and β / hr, respectively
Fine-tuning LLM (Section 4.2.1)
Refining LLM Predictions (Section 4.2.2)
SQL
func()
select
Reasoning Formulation
Rules: : R1. DNA$_A$ -> mRNA$_A$ -> Protein$_A$
R2. [Protein$_A$] = β*[α*[DNA$_A$]]*t
Query: Gene A oscillates periodically.
R2: [Protein$_A$] = β*[α*[DNA$_A$]] * t
[Protein$_A$]$_{MAX}$ = H
Query: Gene A peaks at level Y.
Rule reasoning feedback
Factual knowledge verification feedback
Gene A speed > 0 (0< t <3)
Query: Gene A oscillates periodically. ✗
[DNA$_A$]$_{MAX}$ = H
Query: Gene A peaks at Y. ✗

Figure 7: Our fine-tuning framework.

Overview of the Approach. To further guarantee that LLMs encode **domain-specific biological knowledge** and they are accurate and **grounded to biological principles** (RG1.2), we propose the framework FAIGen for collecting feedback for LLM predictions and fine-tuning. To incorporate scientific principles and domain knowledge into LLM reasoning, and enable them to generalize across contexts, we propose (a) reinforcement fine-tuning from **fine-grained feedback** [42] over biology domain-specific factual and logical errors. (b) fine-tuning by **augmenting synthetic data** generated for unseen situations. We demonstrate the fine-tuning process in Section 5.2.2.

First, we need to collect the feedback for fine-tuning (above) and refinement. We propose FAIGEN which integrates symbolic and neural reasoning for claim verification. We propose **dynamic reasoning tools selection** to verify LLM predictions and provide fine-grained feedback (facts, rules) (bottom of Figure 7). We propose **symbolic representation parsing and execution** to accurately represent conditions/rules and obtain feedback deterministically, thereby addressing faithfulness and generalizability issues. We demonstrate the feedback collection and prediction refinement process in Section 5.2.1.

5.2.1 Feedback Collection and LLMs Predictions Refinement with Scientific/Domain Principles

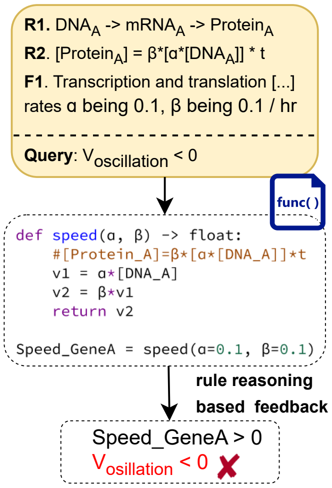
Given the claim (from LLM prediction) and corresponding extracted rules and facts, our goal is to design techniques for automatic verification and collecting feedback, then conduct refinement of the claim to make it grounded to the scientific knowledge, and use the feedback for finetuning (Section 5.2.2).

We propose a neural-symbolic claim verification approach, drawing insights from our work on the LLM tool use [74, 95]. It utilizes LLM to generate the verification plan consisting of multiple modules of tool use (Figure 8). Given the claim/query (from LLM predictions) and corresponding extracted principles, we use GPT-4 to generate a plan of the sequence of reasonings modules (based on rules) and what tool to use for each, then run **deterministic executions** to obtain the **fine-grained feedback/results** (i.e. whether the claim is grounded to *each type* of scientific principles) with **faithfulness guarantee**.

Based on our requirements on the three aspects: scientific **rules**, scientific **facts**, and **logical constraints**, we will design a comprehensive tool set. More specifically, (a) for scientific rules involving quantities, we will use Python interpreter and Wolfram Alpha [95]; (b) for scientific rules of logical relationships and constraints, we will use first-order logic solver and logical programming solver [71]; (c) for relation-based reasoning over scientific facts/knowledge, we will use a knowledge graph QA system (e.g. SPARQL).

Tool Selection and Execution for Faithful Claim Verification. Extending from our neural-symbolic logical reasoning and parsing work [29], for each reasoning module we *parse the principle and query into symbolic representations* based on the tool's documentation. For example, the gene speed calculation module parses the rule into Python function (def speed(α, β)) and the fact of α, β values into statements. Afterward, the framework runs the tools solvers in a deterministic way, which guarantees the faithfulness of the process. The neural-symbolic formulation can more accurately represent more complex situations (e.g. compositional and unseen). It helps **generalization of rules across contexts**, especially previously unobserved situations [59].

Self-refinement from Fine-grained Feedback. We translate the execution results (verification feedback) into natural language feedback including correctness and explanations against scientific principles. Extending from our work on iterative refinement of hypotheses [93], we prompt the base LLM with the feedback to refine the predicted claims (upper part of Figure 2). Benefiting from our framework's comprehensive verification, we obtain *fine-grained feedback* [88] for *all unit claims*, categorized into *three specific categories*: rules, facts, and logical constraints. The refinement process can be iterative until the system determines that no further refinement is necessary. **Risks and Mitigations.** To speed up the refinement, we set the maximum number of iterations to four [58]. **Evaluation Plan.** We will evaluate the refiner in the synthetic biology gene sequence prediction scenario, with prompts designed by PhD students in biomedical engineering. To measure how faithful the predictions are to scientific principles, we will apply our proposed verification stages to verify the correctness of LLM-generated predictions.


R1. DNA$_A$ -> mRNA$_A$ -> Protein$_A$
R2. [Protein$_A$] = β*[α*[DNA$_A$]] * t
F1. Transcription and translation [...] rates α being 0.1, β being 0.1 / hr
Query: V$_{oscillation}$ < 0
func( )
rule reasoning based feedback
Speed_GeneA > 0
V$_{osillation}$ < 0 ✗

```
def speed(α, β) -> float:
    #[Protein_A]=β*[α*[DNA_A]]*t
    v1 = α*[DNA_A]
    v2 = β*v1
    return v2

Speed_GeneA = speed(α=0.1, β=0.1)
```

Figure 8: Planning and Execution.

5.2.2 Fine-tuning LLM for Faithful Biology Knowledge Acquisition and Reasoning

Apart from refining LLM predictions to be grounded to scientific laws (introduced in RG1.2), more importantly, we propose incorporating scientific principles into the **training (fine-tuning) procedures of LLMs** using the collected feedback, to improve their reasoning capabilities and domain expertise/knowledge.

Reinforcement Learning with Fine-grained Rewards. We collect a large set of diverse biology discipline questions to prompt the LLMs to generate content. Moreover, to enable the model to learn to adapt to abnormal unseen situations in the pre-training, we will specifically design more challenging adversarial prompts, such as "Gene A peaks at close to infinite". We use the verification feedback collected in Section 5.2.1. The fine-grained feedback is used to fine-tune the LLM. We draw insights from our work on hallucination mitigation via reinforcement learning with fine-grained feedback [41]. We design rewards R_{rule} and R_{fact} to assess rule-based reasoning errors and factual accuracy. Finally, we design the weighted rewards $R = \alpha R_{rule} + (1-\alpha) R_{fact}$. LLM is updated with R and Proximal Policy Optimization (PPO) [70], **Risks and Mitigations.** To prevent the model from catastrophic problems, we will use a smaller learning rate and smaller training epoch, based on our experience [41].

Fine-tuning with Simulation-based Synthetic Data. It is challenging to ensure LLM's faithful reasoning generalizes across situations (e.g. gene regulations), due to the lack of training data for unobserved settings, especially for scientific disciplines such as biology.

Drawing insights from data synthesis in semantic parsing [38], we propose **generating synthetic data for unobserved conditions for additional fine-tuning**. Specifically, (1) we construct unseen simulated conditions including objects, relations, and attribute values (e.g., "transcription rate=high"); then (2) we use them to initialize the variables in the symbolic form of the scientific principles (e.g., Python program modeling mass-action system equations). Optionally, we conduct concatenations and compositions [38] of principles for complexity. Finally, we translate logic-form to natural language with LLM. For example, in the synthetic biology domain, we might simulate conditions where each gene in a network is transcribed at different rates (e.g., "gene A transcription rate=5 molecules/second") and observe the resulting oscillation patterns. We initialize these variables in a model based on the central dogma of molecular biology

and mass-action kinetics. By combining different rates and conditions (e.g., "transcription rate, translation rate, degradation rate"), we create complex scenarios (e.g., "oscillation rate of gene expression in response to varying transcription factors") and translate these logical representations into natural language descriptions for the LLM to process and learn from. We will fine-tune LLM with the synthetic data to help it conduct more faithful scientific knowledge-guided reasoning, under various including unseen situations.
Risk Mitigations. In case the generated large amount of synthetic data is of bad quality causing a waste of time on training, we will start with a small dataset of more diversity and higher quality (1k) [97] for instruction tuning, then increase to 10k later. **Evaluation Plan.** To measure how faithful the predictions are to scientific principles, we will apply our proposed verification stages in Section 5.2.1, we will compare the performance with the LLM without fine-tuning. For the reward model, we will log the training loss curve and check how it changes over epochs. For synthetic data generation, we will evaluate the quality with human evaluations, also how they contribute to the faithful predictions.

6 RG3: Knowledge Graph-augmented LLM agent for Research Hypotheses Proposal and Experiment Design

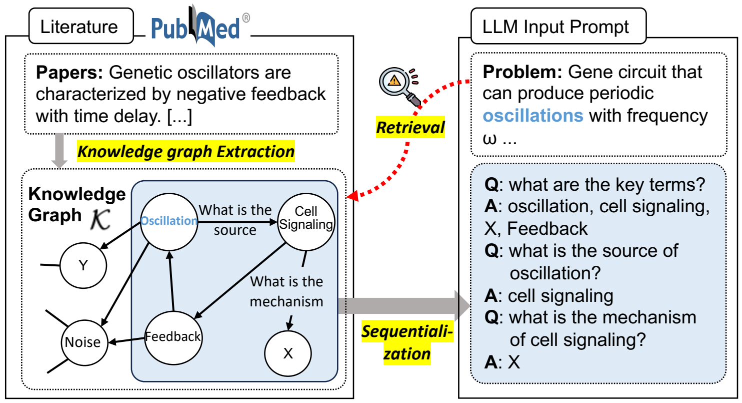
Task Overview and Challenges. Given a problem $\boldsymbol{p}$ (e.g. "gene circuit that can produce [...]" in Figure 9), our main goal is to generate a hypothesis (including the circuit representation and/or properties) and corresponding experimental design. To accomplish the aforementioned steps, the existing literature (e.g., PubMed) is used as a primary source, which provides insights about existing knowledge along with gaps and unanswered questions. Formally, let $\mathcal{L}$ be the literature, and $\boldsymbol{o}$ be the ideas that consist of the hypothesis $\boldsymbol{h}$, and experiment design $\boldsymbol{d}$, as follows: $\boldsymbol{o} = [\boldsymbol{h}, \boldsymbol{d}]$ where each item consists of a sequence of tokens and $[\cdot]$ denotes a concatenation operation. Then the research idea generation process can be represented as $\boldsymbol{o} = LLM(\boldsymbol{p}, \mathcal{L})$. We use the LLM trained in RG2 (Section 5).

One major challenge is that the literature is massive, due to the constraints of their input lengths and their reasoning abilities, particularly over long contexts [57], it is not possible to incorporate all the existing publications from the literature $\mathcal{L}$ into the LLM input.

Based on our expertise in natural language processing, we propose using LLM for extracting entities and their relations from the literature (IE) [22], to construct a knowledge graph (KG), which encodes key knowledge and is structured. The structure would include entities of certain roles (e.g. NAME, MECHANISM, OUTPUT) and will be utilized for downstream LLM predictions. For example, "AA is a genetic oscillator consisting of three genes that inhibit each other. The oscillators are characterized by a negative feedback loop that results in cell signaling[...]" The structure {**Entity (Nodes)**: AA (role PARTICIPANT), cell signaling (RESULT); **Relations (Edges)**: cell signaling-oscilations (SOURCE-OF), genes-AA (PART-OF)} is to be extrated. However traditional general domain IE faces challenges: 1) limited coverage of entities: current methods only extract entities of a pre-defined set of roles, which is coarse-grained and cannot represent the context comprehensively [52], e.g. fail to extract the relation gene-gene of type AGAINST. 2) lack of generalization and data efficiency: models are trained on annotated data and are hard to generalize to new domains and predicate roles [17]; 3) lack of runtime efficiency of the extraction process: current methods extract predicates of each role in separate passes [10, 32]. For example, they first conduct extraction of entities of PARTICIPANT and then for RESULT, etc. It is of high complexity and cost. To address the challenges in extracting structured knowledge from unstructured literature, we propose an integrated approach that leverages advances in large language models (LLMs) for improved generalizability and computational efficiency. Our approach consists of two key strategies: i) We introduce a **question-answer pair representation** and generation-based paradigm to comprehensively extract predicates and improve runtime efficiency; ii) during the generation, we utilize LLMs for zero-shot in-context learning and synthetic data generation to enhance generalization across domains. Generally, we aim to develop knowledge extraction systems that are both generalizable and computationally efficient.

6.1 Extracting Comprehensive Knowledge from Vast Scientific Literature

Our QA-based structured knowledge extraction framework introduces a novel paradigm for **representing predicates by generating question-answer pairs (QAG)** (left bottom of Figure 9). Our goal is to ask questions about the objects in the sentence comprehensively. Under this QAG paradigm, we conduct relation extraction, a set of questions is asked about the objects, whose answer must be another one in the sentence. The questions represent various relations. Different from the close-schema paradigm, where relations within a fixed set of types (e.g. CAUSE-OF and PART-OF) are extracted. Our QA paradigm covers a comprehensive and more nuanced set of types (e.g. "What is the MECHAISM of"). It enables the downstream reasoner to delve deeper into the content.


Literature
PubMed
Papers: Genetic oscillators are characterized by negative feedback with time delay. [...]
Knowledge graph Extraction
Knowledge Graph $\mathcal{K}$
Oscillation
What is the source
Cell Signaling
What is the mechanism
Y
Noise
Feedback
X
Retrieval
Sequentialization
LLM Input Prompt

Problem: Gene circuit that can produce periodic oscillations with frequency ω ...

Q: what are the key terms?
A: oscillation, cell signaling, X, Feedback
Q: what is the source of oscillation?
A: cell signaling
Q: what is the mechanism of cell signaling?
A: X

Figure 9: Leveraging Sequatialized Knowledge Graph (KG) for Research Idea Generation.

We propose a **large language model-based (LLM) generative method** for generating all possible question-answer pairs (entity-relation triples) of the document in one pass across types, instead of one at a time in isolation [21]. We use the following output format: Question$_1$ Answer$_1$ `[SEP]` Q$_2$ A$_2$ `[SEP]`...Q$_i$ A$_i$ `[END]`. This will significantly improve the runtime efficiency. Plus, our goal includes reducing the cost of learning/building extraction systems while still enabling them to generalize across domains. We propose LLM zero-shot prompting which directly outputs the QA sequence and doesn't involve training.

Although LLM excels at various language generation tasks, basic LLM prompting-based information extraction systems (with default prompt) generally under-perform traditional supervised IE systems, especially on specific domains/settings [51, 91]. Challenges include forcing the correct format and validness of generated predicate question-answer pairs, we design novel **in-context learning prompting** strategies for the extractions. Specifically, to better elicit reasoning for following the template while generating the extracted results, we propose leveraging Chain-of-Thought (CoT) [84] style explanations in the prompt. For example, the reasoning explanation for extracting "Q: what is PART-OF AA? Ans: genes." would be "since AA consists of three genes, the relation is PART-OF." Moreover, for small sub-domains where LLM lacks relevant knowledge, we propose **synthetic data generation** [43] for the instruction tuning of LLMs. Generating high-quality synthetic data is challenging, we will conduct *inverse generation*. It prompts LLMs to produce natural text based on the existing structural data provided as input (entity-relation represented by QA pairs). This strategy ensures that the predicate is appropriately displayed within the input text and that the training example is challenging and meaningful. **We denote the extracted knowledge store $\mathcal{K}$.**

6.2 Retrieving and Leveraging Relevant Knowledge for Generating Hypotheses and Experiment Design

Given this knowledge store $\mathcal{K}$, our goal is to enhance the vanilla research idea generation process based on given problem requirements and context, denoted as follows: $\boldsymbol{o} = \texttt{LLM}(\boldsymbol{p})$. We do this by augmenting the `LLM` with the relevant knowledge triples from $\mathcal{K}$, which can expand the contextual knowledge – what LLMs can consume – by offering additional knowledge. In other words, this knowledge is not seen in the original input but is relevant to it. More specifically, we retrieve the top m triples (question answer pairs) from the knowledge graph with DPR [46]. After that, we append the retrieved knowledge to the input problem $\boldsymbol{p}$ in the form of the prompt (right part of Figure 9), which is then forwarded to LLMs to generate the research idea. Our framework, Knowledge-Augmented LLM, requires no model training. Hereafter, the instantiation of research proposal generation augmented with relevant entity-centric knowledge is represented as follows: $\boldsymbol{o} = \texttt{LLM}(\boldsymbol{p}; \mathcal{K})$. We call this knowledge-augmented LLM-powered idea generation agent.

7 RG4: Refining Hypotheses with Conceptual & Experiment Feedback

Overview and Motivations. The candidate hypotheses generated by the one-step generation approach (that concludes once the ideas are formulated) in Section 6 are of low quality and lack a refinement process based on reviews, which differs from typical human-driven research processes, which develop and improve research ideas through inspecting the hypotheses from multiple perspectives; and conducting experiments for validations. To tackle those limitations, we further propose to (1) expand the idea generation process by iteratively refining the generated ideas through feedback from LLM reviewing agents (Section 7.1); (2) conduct experiments following the generated experimental plan to provide feedback (Section 7.2).

7.1 RG4.1: Hypotheses Refinement via Conceptual Feedback

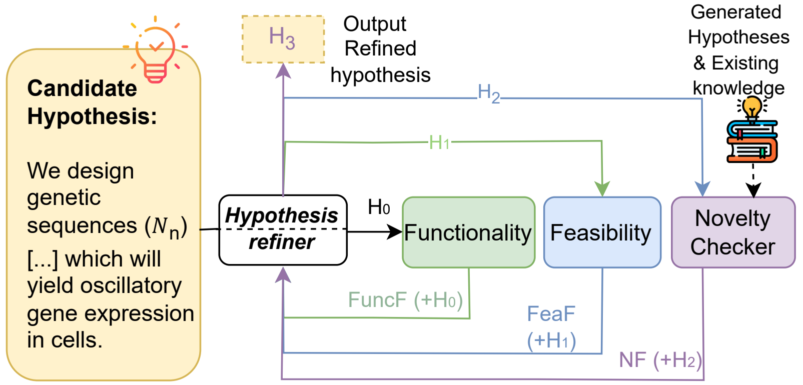

Candidate Hypothesis:
We design genetic sequences (N_n) [...] which will yield oscillatory gene expression in cells.
H_3
Output Refined hypothesis
H_2
H_1
Generated Hypotheses & Existing knowledge
Hypothesis refiner
H_0
Functionality
Feasibility
Novelty Checker
FuncF (+H_0)
FeaF (+H_1)
NF (+H_2)

Figure 10: Multi-module refinement framework with inter-module iterative feedback. Each hypothesis is passed through Feasibility, Functionality, and Novelty checkers for feedback and then refined.

We propose the multi-module hypothesis refinement process as shown in Figure 10. Specifically, the refinement follows an *iterative process*. Inspired by philosophy research on inductive reasoning [66], apart from the main hypothesis refiner module (Ref), we will design checker modules. Based on our domain knowledge on synthetic biology, we determine three key aspects of refining/evaluating the generated circuits, i.e. feasibility, functionality, and novelty. For each of the aspect, one checker module is designed. Together, they provide feedback for the generated hypothesis sequentially. Each checker module can be initialized with an LLM, which will be capable of providing feedback. For each candidate hypothesis H_0 induced from the proposer, we prompt the functionality checker with the input consisting of H_0 and instruction ‘‘`Given a research hypothesis:{hypothesis}, determining if the genetic circuit performs the desired biological function effectively, such as oscillation, signal amplification.`’’ to obtain the functionality feedback (FF), where `{hypothesis}` is initialized with the candidate hypothesis. Then hypothesis refiner Ref takes both H_0 and $FuncF$ as input and utilizes the feedback to generate the refined hypothesis H_1; the process can be expressed as $H_1 = Ref(H_0, FuncF)$. Similarly, the feasibility checker generates feedback $FeaF$ based on H_1 and the instruction ‘‘`Given a research hypothesis:{hypothesis}, give feedback on whether the proposed genetic circuit can be realistically constructed and implemented using current synthetic biology techniques.`’’ [66]. This feedback is later input to the refiner: $H_2 = L(H_1, FeaF)$. Next, we obtain the final hypotheses via novelty feedback NF: $H_3 = Ref(H_2, NF)$. For example, the feedback on novelty might be “The research hypothesis does not offer a significantly novel perspective, since it appears in existing literature [1,2]. A more novel idea would be leveraging [3] to ...”. The *multi-step nature of our framework contributes to the faithfulness of the resulting hypotheses* since a succeeding step relies on the previous step’s generation, and the explanations are generated in a sequential manner and more fine-grained (for different aspects), which is different from end-to-end generation setting of Chain-of-Thought [84].

Different from functionality and feasibility checking, the novelty checking phase is closely related to real-world knowledge and requires more faithful and grounded feedback [31]. As a result, we will develop a novel specialized retrieval system (rightmost of Figure 10) to retrieve from the hypotheses store consisting of (1) the already generated hypotheses from our framework; and (2) the existing important findings or knowledge. Then the candidate hypothesis and retrieved contents are fed into the novelty checker.

7.2 RG4.2: Hypotheses Refinement via Experimental Feedback

Our hypothesis generation technique will be supported by an experimental setup designed to thoroughly evaluate a wide variety of circuits. **Our primary objectives are to probe the dynamics of the synthetic circuits in the generated hypothesis, using protein expression measurement through time-lapse microscopy over several days, and single-molecule mRNA and single-cell protein measurements as endpoint assays.** To evaluate the robustness of the circuits generated by the hypothesis, synthesized genetic circuits will be introduced to cells in various doses and perturbed from their steady state to study their dynamic response and steady state properties. We expect to obtain rich data that will enable us to quantify key parameters underlying the behavior of synthetic circuits. Theoretical models and statistical methods from the literature will guide the calculation of parameters like expression level, duration, periodicity, and noise. For example, fluorescence intensity data will help derive expression levels and durations, while Fourier transform techniques will enable analysis of periodicity. Statistical measures such as coefficient of variation and Fano factor will be used to quantify noise in expression. These calculated parameters will then be compared directly to the initial conditions as outlined by the hypothesis.

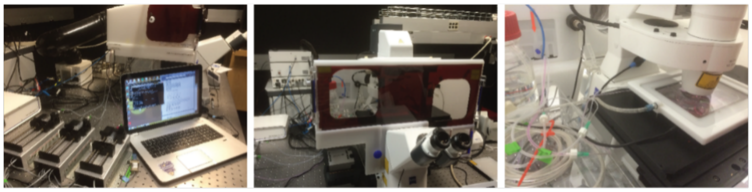


Figure 11: Experimental setup for time-lapse experiments. Custom pumps for long term time-lapse microscopy and automated delivery of reagents.

We will use a custom-fluidic setup (Figure 11), recently established in the Bleris lab, that allows us to perform controlled perturbation and media exchange experiments for prolonged periods of time (i.e. time-lapse experiments lasting 7 days). This custom setup will enable fluorescence time-lapse microscopy with high temporal resolution. For end-point measurement of protein, we will perform flow cytometry. For end-point measurements of mRNA, we will track individual mRNA in cells with PrimeFlow (Affymetrix) technology, a proprietary oligonucleotide probe set design and branched DNA (bDNA) signal amplification method to analyze RNA transcripts by flow cytometry. This technology will allow us to rapidly assess mRNA distributions under different experimental conditions. For example, we will deactivate the fluorescent proteins (introducing mutations) and probe the three nodes in parallel at the mRNA level. Alternatively, we will quantify each node separately and obtain distributions that map the mRNA and protein abundances for each node.

Evaluation Plan

We will evaluate the proposed framework on our dataset, which contains a curated list of 100 commonly observed gene network motifs, such as one in Figure 2. Each prompt will be accompanied by manually annotated information, including the optimal hypothesis and research background. For the evaluation of induced hypothesis circuits, we will use the criteria of functionality, novelty, and robustness using a 1-5 Likert scale, and perform an expert manual evaluation and automatic evaluation. The functionality score will be given based on how well the genetic circuit's measurable parameters (e.g. frequency of oscillation) align with the desired characteristics. The novelty score will reflect the uniqueness of the genetic circuit's topology compared to the expected topology in the existing literature. Robustness will be determined by the genetic circuit's ability to consistently produce the desired output despite variability in factors such as cell type, copy number variation, and noise. **Expert manual evaluation** will be performed by graduate students in the Bioengineering Department of UT Dallas. For **automatic evaluation**, we will utilize GPT-4 for generating the scores. In the prompt, we will include background information such as the knowledge store and experimental results. We will also use the above methods to evaluate the intermediate generations and explanations as to whether they satisfy the goals for each feedback.

8 Project Schedule

The proposed project is expected to take 24 months to finish and will involve two graduate students each year. The detailed research goals and timeline are shown in the Figure.

Research Goals	**Year1**				**Year2**			
RG1: Synthetic Biology Task Preparation, Domain Knowledge and Dataset Collection (PI Bleris)	🟦	🟦	🟦	🟦				
RG2: Biology-inspired Training of Large Language Models								
Evolutionary-inspired Model Training and Merging (PI Du and Bleris)			🟦	🟦	🟦	🟦		
Biology Knowledge-guided Fine-tuning (PI Du and Bleris)					🟦	🟦	🟦	🟦
RG3: Knowledge Graph-augmented LLM agent for Research Idea Generation								
Extracting Comprehensive Structured Knowledge from Scientific Literature (PI Du)	🟦	🟦	🟦	🟦				
Retrieving and Leveraging Relevant Knowledge for Generating Research Idea (PI Du)			🟦	🟦	🟦	🟦		
RG4: Refining Hypotheses with Conceptual and Experimental Feedback								
Hypotheses Refinement via Conceptual Feedback (PI Du)				🟦	🟦	🟦	🟦	🟦
Hypotheses Refinement via Experimental Feedback (PI Bleris)				🟦	🟦	🟦	🟦	🟦
Educational Goals (PI Du and Bleris)	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦

9 Results from Prior NSF Support

PI Leo Bleris. Intellectual merit. Bleris received support from NSF to study microRNA sensors (Grant #1105524, $336k total, 09/01/11-8/31/14, "Detecting Cancer at the Single-Cell Level Using Endogenous Signal Biomolecular Sensors"). Published work from the grant includes: transcription activator-like effector hybrids for microRNA-based control of stably integrated genes [56], a range of synthetic intragenic miRNAs co-expressed with their host genes [47], and a study on the effect of negative feedback on global and local sources of uncertainty in mammalian transgenes [78]. Bleris received an NSF CAREER award to study TALE libraries (Grant #1351354, $400k total, 09/01/14-8/31/19, "CAREER: Versatile transcription activator-like effector libraries for genome-wide screens"). TALEs are a class of specific DNA binding proteins. We developed protocols for constructing custom TALE proteins for specific nucleotide sequence binding [56], a novel approach to construct versatile transcription-activator like effector libraries [54, 62] (IP pending), and a novel biological decoder [27]. Bleris received (as co-PI) a grant from a DMS/NIGMS program (Grant #1361355, $744k total, 09/01/14-8/31/18, "Single-cell to population-based analysis of the microRNA-p53 interaction network") to study the miRNA-MDM2-p53 network. We recently showed that miR-192-mediated positive feedback controls the robustness of stress-induced p53 oscillations in breast cancer cells [62] and published genome editing tools and analysis [55, 67].

Figure 12: Left: High school students participating in the Texas State Science Fair and receiving the First Grand prize. Middle/Right: Bleris lab members at Perot Science Museum interacting with the public during the "Building with Biology" event.

Broader impacts In terms of the broader impacts (from Grants #1105524 and #1351354), the PI Bleris mentored undergraduate students (e.g.[62, 63, 64] second author undergraduate and [45] first author undergraduate). The results of sponsored research work were directly integrated in the class "Systems Biology". In terms of **wider dissemination**, the PI Bleris assists in organizing workshops and sessions in conferences and actively presents research results as an invited speaker. In terms of the broader impacts, in order to **integrate science and engineering with K-12 education**, our group delivered a lesson in biology targeted at various age groups at a local summer camp for at-risk youth. Specifically, we developed a lesson plan for first graders. The Bleris lab has offered **summer research opportunities to high school students**. Yesh Doctor and Kshitij Sachan were supported and participated in several science competitions that include: (a) Plano East Science Fair - First place, (b) Texas State Science and Engineering Fair - First grand prize (life sciences) (Figure 12), (c) International Sustainable World - gold medal, (d) Siemens Competition in Math Science and Technology - Regional Finalist. Finally, members of the Bleris Lab participated in "Building with biology" at the Perot Museum (Figure 12), an event that creates conversations among scientists and public audiences about the emerging field of synthetic biology and its societal implications.

PI Xinya Du. CAREER: Learning to Extract Consistent Event Graphs from Long and Complex Documents. Du, X. (2024-2029), Award Number: 2340435, $561,219.00.

Intellectual merit. This project aims to break new ground in the area of event knowledge acquisition from long documents. To now, we have investigated the new paradigm of using question-answer pairs for event extraction. We constructed and made available one dataset [9], and are planning to employ them in the later part of the project. We are developing methods for improving the efficiency of EKG construction.

Broader impacts. Up to now, the project provided graduate training and mentoring to one PhD student Ruochen Li, who submitted a paper to CIKM. It provides extensive training to one master student Milind Choudhary, who recently had a paper accepted at EACL [9] and was admitted to UTD CS PhD program. It provides extensive training to one undergraduate research assistant Teerth Patel, who recently submitted a paper to the journal TMLR [53].

References Cited

[1] UTD Research Program for HS Students, 2024. https://k12.utdallas.edu/research/.

[2] M. R. AI4Science and M. Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *ArXiv*, abs/2311.07361, 2023.

[3] T. Akiba, M. Shing, Y. Tang, Q. Sun, and D. Ha. Evolutionary optimization of model merging recipes. *arXiv preprint arXiv:2403.13187*, 2024.

[4] E. Andrianantoandro, S. Basu, D. K. Karig, and R. Weiss. Synthetic biology: new engineering rules for an emerging discipline. *Molecular systems biology*, 2(1):2006–0028, 2006.

[5] J. Baek, S. K. Jauhar, S. Cucerzan, and S. J. Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

[6] I. Beltagy, K. Lo, and A. Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, 2019.

[7] D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

[8] H. Capel, R. Weiler, M. Dijkstra, R. Vleugels, P. Bloem, and K. A. Feenstra. Proteinglue multi-task benchmark suite for self-supervised protein modeling. *Scientific Reports*, 12(1):16047, 2022.

[9] M. Choudhary and X. Du. Qaevent: Event extraction as question-answer pairs generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2024.

[10] M. Choudhary and X. Du. QAEVENT: Event extraction as question-answer pairs generation. In Y. Graham and M. Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1860–1873, St. Julian's, Malta, Mar. 2024. Association for Computational Linguistics.

[11] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, and R. Nakano. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[12] S. Curran, S. Lansley, and O. Bethell. Hallucination is the last thing you need. *arXiv preprint arXiv:2306.11520*, 2023.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] X. Du and C. Cardie. Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[15] X. Du and C. Cardie. Harvesting paragraph-level question-answer pairs from Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1917, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[16] X. Du and C. Cardie. Document-level event role filler extraction using multi-granularity contextualized encoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8010–8020, Online, July 2020. Association for Computational Linguistics.

[17] X. Du and C. Cardie. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online, Nov. 2020. Association for Computational Linguistics.

[18] X. Du and H. Ji. Retrieval-augmented generative question answering for event argument extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4649–4666, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[19] X. Du, S. Li, and H. Ji. Dynamic global memory for document-level argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5264–5275, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[20] X. Du, A. Rush, and C. Cardie. GRIT: Generative role-filler transformers for document-level event entity extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online, Apr. 2021. Association for Computational Linguistics.

[21] X. Du, J. Shao, and C. Cardie. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[22] X. Du, Z. Zhang, S. Li, Z. Wang, P. Yu, H. Wang, T. Lai, X. Lin, I. Liu, B. Zhou, H. Wen, M. Li, D. Hannan, J. Lei, H. Kim, R. Dror, H. Wang, M. Regan, Q. Zeng, Q. LYU, C. Yu, C. N. Edwards, X. Jin, Y. Jiao, G. Kazeminejad, Z. Wang, C. Callison-Burch, C. Vondrick, M. Bansal, D. Roth, J. Han, S.-F. Chang, M. Palmer, and H. Ji. Resin-11: Schema-guided event prediction for 11 newsworthy scenarios. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, 2022.

[23] K. Ehrhardt, M. Guinn, T. Quarton, M. Zhang, and L. Bleris. Reconfigurable hybrid interface for molecular marker diagnostics and in-situ reporting. *Biosensors Bioelectronics*, 74:744–750, 2015.

[24] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rihawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing, 2021.

[25] D. Endy. Foundations for engineering biology. *Nature*, 438(7067):449–453, 2005.

[26] M. Fire and C. Guestrin. Over-optimization of academic publishing metrics: observing goodhart's law in action. *GigaScience*, 8(6):giz053, 2019.

[27] M. Guinn and L. Bleris. Biological 2-input decoder circuit in human cells. *ACS Synthetic Biology*, 3(8):627–633, 2014.

[28] C. Han, Q. He, C. Yu, X. Du, H. Tong, and H. Ji. Logical entity representation in knowledge-graphs for differentiable rule learning. In *ICLR*, 2023.

[29] C. Han, H. Pei, X. Du, and H. Ji. Zero-shot classification by logical reasoning on natural language explanations. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8967–8981, Toronto, Canada, July 2023. Association for Computational Linguistics.

[30] N. Hansen. The cma evolution strategy: a comparing review. *Towards a new evolutionary computation: Advances in the estimation of distribution algorithms*, pages 75–102, 2006.

[31] H. He, H. Zhang, and D. Roth. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*, 2022.

[32] W. Held, D. Iter, and D. Jurafsky. Focus on what matters: Applying discourse coherence theory to cross document coreference. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1406–1417, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[33] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.

[34] T. Hope, D. Downey, D. S. Weld, O. Etzioni, and E. Horvitz. A computational inflection for scientific discovery. *Communications of the ACM*, 66(8):62–73, 2023.

[35] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

[36] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[37] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[38] R. Jia and P. Liang. Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, 2016.

[39] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023.

[40] L. Jing and X. Du. Fgaif: Aligning large vision-language models with fine-grained ai feedback, 2024.

[41] L. Jing and X. Du. Mitigating hallucinations in vision-language models with fine-grained ai feedback, 2024.

[42] L. Jing, R. Li, Y. Chen, M. Jia, and X. Du. Faithscore: Evaluating hallucinations in large vision-language models, 2023.

[43] M. Josifoski, M. Sakota, M. Peyrard, and R. West. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574, 2023.

[44] T. Kang, T. Quarton, C. Nowak, K. Ehrhardt, A. Singh, Y. Li, and L. Bleris. Robust filtering and noise suppression in intragenic mirna-mediated host regulation. *iScience*, 23(10):101595, 2020.

[45] T. Kang, J. White, Z. Xie, Y. Benenson, E. Sontag, and L. Bleris. Reverse engineering validation using a benchmark synthetic gene circuit in human cells. *ACS Synthetic Biology*, 2(5):255–262, 2013.

[46] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

[47] N. Kashyap, B. Pham, Z. Xie, and L. Bleris. Transcripts for combined synthetic microrna and gene delivery. *Molecular BioSystems*, 9(7):1919–1925, 2013.

[48] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024.

[49] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[50] M. Leisner, L. Bleris, J. Lohmueller, Z. Xie, and Y. Benenson. Rationally designed logic integration of regulatory signals in mammalian cells. *Nature Nanotechnology*, 5(9):666–670, 2010.

[51] B. Li, G. Fang, Y. Yang, Q. Wang, W. Ye, W. Zhao, and S. Zhang. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *ArXiv*, abs/2304.11633, 2023.

[52] Q. Li, H. Ji, and L. Huang. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics.

[53] R. Li, T. Patel, and X. Du. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*, 2023.

[54] Y. Li, K. Ehrhardt, M. Zhang, and L. Bleris. Assembly and validation of versatile transcription activator-like effector libraries. *Scientific Reports*, 4, 2014.

[55] Y. Li, S. Mendiratta, K. Ehrhardt, N. Kashyap, M. White, and L. Bleris. Exploiting the crispr/cas9 pam constraint for single-nucleotide resolution interventions. *PLoS ONE*, 11(1):e0144970, 2016.

[56] Y. Li, R. Moore, M. Guinn, and L. Bleris. Transcription activator-like effector hybrids for conditional control and rewiring of chromosomal transgene expression. *Scientific Reports*, 2, 2012.

[57] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2023.

[58] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[59] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. International Conference on Learning Representations, 2019.

[60] V. Marx. The big challenges of big data. *Nature*, 498(7453):255–260, 2013.

[61] M. S. Matena and C. A. Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

[62] R. Moore, A. Chandrahas, and L. Bleris. Transcription activator-like effectors: a toolkit for synthetic biology. *ACS Synthetic Biology*, 3(10):708–716, 2014.

[63] R. Moore, H. Ooi, T. Kang, L. Bleris, and L. Ma. microrna-192-mediated positive feedback loop controls the robustness of stress-induced p53 oscillations in breast cancer cells. *PLoS Computational Biology*, 11(12):e1004653, 2015.

[64] R. Moore, A. Spinhirne, M. Lai, S. Preisser, Y. Li, T. Kang, and L. Bleris. Crispr-based self-cleaving mechanism for controllable gene delivery in human cells. *Nucleic Acids Research*, 43(2):1297–1303, 2015.

[65] E. Nguyen, M. Poli, M. Faizi, A. W. Thomas, M. Wornow, C. Birch-Sykes, S. Massaroli, A. Patel, C. M. Rabideau, Y. Bengio, S. Ermon, C. Ré, and S. Baccus. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[66] J. D. Norton. A little survey of induction. 2003.

[67] C. Nowak, S. Lawson, M. Zerez, and L. Bleris. Guide rna engineering for versatile cas9 functionality. *Nucleic Acids Research*, 2016.

[68] OpenAI. Gpt-4 technical report, 2023.

[69] OpenAI. Gpt-4v(ision) system card, 2024. https://cdn.openai.com/papers/GPTV_System_Card.pdf.

[70] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[71] L. Pan, A. Albalak, X. Wang, and W. Y. Wang. Logic-LM: empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP)*, Singapore, Dec 2023.

[72] P. E. Purnick and R. Weiss. The second wave of synthetic biology: from modules to systems. *Nature reviews Molecular cell biology*, 10(6):410–422, 2009.

[73] B. Qi, K. Zhang, H. Li, K. Tian, S. Zeng, Z.-R. Chen, and B. Zhou. Large language models are zero shot hypothesis proposers. *arXiv preprint arXiv:2311.05965*, 2023.

[74] C. Qian, C. Han, Y. Fung, Y. Qin, Z. Liu, and H. Ji. Creator: Tool creation for disentangling abstract and concrete reasoning of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6922–6939, 2023.

[75] T. Quarton, K. Ehrhardt, J. Lee, S. Kannan, Y. Li, L. Ma, and L. Bleris. Mapping the operational landscape of micrornas in synthetic gene circuits. *npj Systems Biology and Applications*, 4(1):1–7, 2018.

[76] T. Quarton, T. Kang, V. Papakis, K. Nguyen, C. Nowak, Y. Li, and L. Bleris. Uncoupling gene expression noise along the central dogma using genome engineered human cell lines. *Nucleic Acids Research*, 48(16):9406–9413, 2020.

[77] K. Rinaudo, L. Bleris, R. Maddamsetti, S. Subramanian, R. Weiss, and Y. Benenson. A universal rnai-based logic evaluator that operates in mammalian cells. *Nature Biotechnology*, 25(7):795–801, 2007.

[78] V. Shimoga, J. White, Y. Li, E. Sontag, and L. Bleris. Synthetic mammalian transgene negative autoregulation. *Molecular Systems Biology*, 9(1), 2013.

[79] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pages 9120–9132. PMLR, 2020.

[80] Y. Tang, Y. Tian, and D. Ha. Evojax: Hardware-accelerated neuroevolution. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 308–311, 2022.

[81] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[82] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, and A. Deac. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[83] Q. Wang, D. Downey, H. Ji, and T. Hope. Scimon: Scientific inspiration machines optimized for novelty. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

[84] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.

[85] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019.

[86] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. G. Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, and L. Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR, 2022.

[87] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7949–7961. IEEE, 2022.

[88] Z. Wu, Y. Hu, W. Shi, N. Dziri, A. Suhr, P. Ammanabrolu, N. A. Smith, M. Ostendorf, and H. Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *arXiv preprint arXiv:2306.01693*, 2023.

[89] Z. Xie, S. Liu, L. Bleris, and Y. Benenson. Logic integration of mrna signals by an rnai-based molecular computer. *Nucleic Acids Research*, 38(8):2692–2701, 2010.

[90] Z. Xie, L. Wroblewska, L. Prochazka, R. Weiss, and Y. Benenson. Multi-input rnai-based logic circuit for identification of specific cancer cells. *Science*, 333(6047):1307–1311, 2011.

[91] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, and E. Chen. Large language models for generative information extraction: A survey, 2023.

[92] Z. Yang, L. Dong, X. Du, H. Cheng, E. Cambria, X. Liu, J. Gao, and F. Wei. Language models as inductive reasoners. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, Malta, 2024. Association for Computational Linguistics.

[93] Z. Yang, X. Du, J. Li, J. Zheng, S. Poria, and E. Cambria. Large language models for automated open-domain scientific hypotheses discovery, 2023.

[94] Z. Yang, X. Du, R. Mao, J. Ni, and E. Cambria. Logical reasoning over natural language as knowledge representation: A survey. *arXiv preprint arXiv:2303.12023*, 2023.

[95] L. Yuan, Y. Chen, X. Wang, Y. R. Fung, H. Peng, and H. Ji. Craft: Customizing llms by creating and retrieving from specialized toolsets. *arXiv preprint arXiv:2309.17428*, 2023.

[96] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[97] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy. LIMA: less is more for alignment. *CoRR*, abs/2305.11206, 2023.